SDC-0055-051-LF-3019A
June 10, 1994

RULES FOR REPRESENTATION OF CHEMICAL DATA IN GATEWAY/ENVIROFACTS PILOT MASTER CHEMICAL INTEGRATOR

CONTRACT NO. 68-W1-0055
DELIVERY ORDER NO. 051

Prepared for:

United States Environmental Protection Agency
Office of Information Resources Management
401 M. Street, SW.
Washington, DC 20460:

Delivery Order Project Officer:

William L. Muldrow

Prepared by:

EPA Systems Development Center
(A Contractor Operated Facility)
Science Applications International Corporation
200 North Glebe Road, Suite 300
Arlington, VA 22203

CONTENTS

EXECUTIVE SUMMARY

Introduction

The U.S. Environmental Protection Agency (EPA) is concerned with the need to provide environmental managers with convenient access to a full range of information necessary for effective, risk-based decision making. The need for integration of EPA's information resources is being addressed by the Office of Information Resource Management (OIRM) through a combination of approaches. This document addresses the integration of EPA's information resources based on chemical identity.

The EPA chemical data standard (EPA 2180.1) establishes the policy and responsibilities related to the use of the Chemical Abstracts Services Registry Number (CASRN) for the identification of specific, definable chemical substances in EPA computer systems. The EPA, however, regulates chemical substances that are not specific, definable chemicals (e.g., chemical categories regulated under Title III of the Emergency Planning and Community Right-to-Know Act and chemical substances of ambiguous composition, such as diesel fuels). In addition, EPA inquiries may require information from EPA environmental systems for a group of related substances, in addition to information about a specific chemical.

This document contains proposed rules for identification of chemical substances across the broad range of EPA concerns to facilitate the integrated recovery of data from EPA information systems. The rules will be used in a pilot chemical integrator module within the GATEWAY/ENVIROFACTS system (i.e., the ENVIROFACTS Master Chemical Integrator (EMCI)). In addition, the rules are intended to provide further guidance to enhance the EPA chemical data standard.

Scope of Proposed Rules

The proposed rules shall be applicable for representing the following kinds of data:

- Specific unique chemical substances.
- Chemical substances with undefined molecular structure.
- Chemicals for which CASRNs have not been assigned.
- Ions, where the complete salt is not known (e.g., Nitrates).
- Different valence states of an ion (e.g., Ferric and Ferrous).
- Radioisotopes (e.g., Thorium^{227} and Thorium^{228}).
- Groups of chemicals (e.g., categories, mixtures, and non-specific substances).

Proposed Rules

The following data are required to be stored for each chemical substance:

- Identification number.
- A chemical name that provides a unique definition of the substance.

The identification number shall be the CASRN, where a number for the substance has been assigned by the Chemical Abstracts Service (CAS) to that specific substance. The CASRN shall be stored as a right-justified, zero-filled, 9-digit number and displayed with hyphens in the CAS display format. The CASRN shall be subjected to the CAS check digit algorithm to ensure that the number is valid. If a CASRN is unavailable, the EMCI software shall generate a unique identification number for that substance.

The most specific, systematic name available shall be used to identify the substance uniquely. The CAS Index Name, stored in inverted format, is the preferred name for identification of a substance. Additional names (i.e., names stored in the ENVIROFACTS component systems) will be stored as synonyms for the substance. Synonyms may be systematic, trivial, or trade names.

EMCI Data Management

Program offices responsible for environmental data systems that constitute the ENVIROFACTS component systems shall provide access to the EMCI data management team for regularly scheduled downloading of environmental data and chemical data validation tables. They shall also provide the EMCI data management team with up-to-date documentation for their systems to ensure that the data relationships are accurately represented in ENVIROFACTS and shall respond to data inconsistencies identified by the ENVIROFACTS support group.

The EMCI chemical data management team shall be responsible for maintaining chemical data validation tables received from the component systems, determine the most appropriate name and the correct CASRN for identification of the substance in the EMCI master file, and update the master chemical index of ENVIROFACTS chemicals and the cross-references to program office systems' chemicals. The data management team will notify the appropriate program office(s) of data errors or inconsistencies that are observed in the component systems.

Continuing Actions

This document is intended to be a living document that will be revised and appended throughout the development of the EMCI to conform to data needs, system constraints and users' needs to search and retrieve data. The document will be further enhanced as needed to meet the requirements for the future development of an Agency-wide Central Chemical Index System.

1.0 INTRODUCTION

The U.S. Environmental Protection Agency (EPA) is concerned with the need to provide environmental managers with convenient access to a full range of information necessary for effective, risk-based decision making. Traditionally, EPA data systems have been developed by program offices to support the management of separate elements of environmental legislation. These systems have been developed on a variety of platforms, both hardware and software, with little regard for the transfer of data across systems. The need for integration of EPA's information resources is being addressed by the Office of Information Resource Management (OIRM) through a combination of approaches. Two approaches towards integration, addressed in this document, are the following:

- Data standards.
- GATEWAY/ENVIROFACTS integrated database system.

1.1 Data Standards

The data standards approach towards addressing data integration has been the development and implementation of data standards and policies to improve Agency-wide data consistency and sharing potential across EPA data systems.

The data standard of concern to this document is the *Chemical Abstracts Service (CAS) Registry Number Data Standard*, issued on June 26, 1987, as EPA 2180.1. The purpose of EPA 2180.1 was to establish the policy and responsibilities related to the use of registry data, specifically the CAS Registry Number (CASRN), from the CAS Division of the American Chemical Society. The CASRN is intended to provide consistent and unambiguous identification of chemicals and to facilitate sharing chemical information across programmatic media. Briefly stated, the policy, according to EPA 2180.1, is:

- Any computer-based Agency system currently in use or being planned containing *data/information on specific, definable chemical substances* shall contain the CASRN for each chemical substance.

- *Additional data selected from the CAS chemical registry system*, such as CAS Index Names and Synonyms, Molecular Formulas and the CAS Chemical Registry Records, *are optional*.

A previous project, the Data Standards Implementation Program (DSIP), analyzed the implementation of EPA data standards in a selected set of EPA environmental systems. The results of the DSIP included recommendations for chemical data management, including development of a master chemical integrator to facilitate chemical data integration across EPA systems. Documents produced under the DSIP project that were used in the preparation of these proposed rules include the *Audit Analysis Report for the Data Standards Implementation Program*, *Recommendations for a Data Standards Implementation Strategy*, and the *Mission Needs Analysis for a Central Chemical Index System*.

1.2 GATEWAY/ENVIROFACTS Integrated Database System

In 1990, EPA began the development of the GATEWAY prototype as a means of integrating data resources from several of EPA's separate databases. In early 1992, the Great Lakes National Program Office (GLNPO) requested support for their evolving data management needs. It was decided that GATEWAY/ENVIROFACTS would serve as the technical architecture for the GLNPO system and that the system should be expanded to address anticipated GLNPO needs. The system now includes data from the Permit Compliance System (PCS), the Toxic Release Inventory (TRIS), the Facility Index System (FINDS), and the Comprehensive Emergency Response, Compensation and Liability Information System (CERCLIS). The Resource Conservation and Recovery Information System (RCRIS) will be incorporated into version 3.0 of ENVIROFACTS.

The wide range of potential users and uses for GATEWAY/ENVIROFACTS make it essential that ENVIROFACTS be completely standards-based. The basic concept is one of providing access to a broad spectrum of well-organized, documented environmental data sources (ENVIROFACTS) through a standard interface (GATEWAY) operating in the widest range of hardware and software environments. System developers expect that ENVIROFACTS will not only serve as the data storage system for GLNPO and other geographic initiatives, but will also be a prototype for standard information engineering and data management practices for EPA's information resource management community.

ENVIROFACTS uses FINDS to integrate facility identification data across EPA's separate environmental systems. Ongoing enhancements to FINDS will provide the additional capability for FINDS to facilitate the integration of spatial data for the systems. However, integration of data related to chemical substances has not been possible with the existing ENVIROFACTS system. Implementation of the chemical data standard has not been consistent within the Agency, resulting in an absence of reliable chemical data linkages across the environmental systems.

The EPA has now determined that a chemical integrator module shall be developed within the GATEWAY/ENVIROFACTS environment, to facilitate integration of EPA environmental systems based on chemical data.

1.3 Purpose

The purpose of this document is to establish the rules for identification and integration of chemicals in the GATEWAY/ENVIROFACTS environment, in compliance with the chemical data standard and in agreement with the DSIP recommendations for chemical data management. The ENVIROFACTS Master Chemical Integrator (EMCI) module to be developed according to these rules is intended to serve as a model for chemical data representation and to facilitate implementation of the chemical data standard throughout the Agency.

1.4 Existing EPA Chemical Index Systems

Two systems have been developed within the EPA that provide information about EPA-regulated and monitored chemicals. The Register of Lists (RoL) was developed by the Office of Policy, Planning, and Evaluation (OPPE) to identify lists of regulated chemicals; the Environmental Monitoring Methods Index (EMMI) System, developed by the Office of Water Regulations and Standards (OWRS) provides information about analytical methods to be used for environmental monitoring studies. Both of these systems use CASRNs, where available, to identify chemical substances. Both systems have assigned internal identification codes to substances for which no CASRNs have been assigned.

Both the RoL and EMMI group chemical substances into categories, so that related substances can be identified. The RoL uses an internal, undisplayed code for linking categories and mixtures with their component parts. The EMMI uses a "Base CASRN" to indicate relationships. For example, EMMI uses the CASRN for Mercury as a "Base CASRN" for Mercury Salts and the CASRN for Nitric Acid as the "Base CASRN for Nitrates. The method used by EMMI does not allow a substance to be linked to more than one "Base CASRN." Mercury Nitrate, therefore, cannot be related to both Mercury and to Nitric Acid.

Neither of the aforementioned existing systems meets the need for integrating environmental data across EPA program systems. The following are true for both systems:

- System software is PC-based, and the software is not transportable to mainframe or mini-computer use.

- The systems do not provide a pointer to EPA automated systems where environmental data are stored.

1.5 References and Definitions

Appendix A contains a list of reference materials used in the preparation of the document. Appendix B contains definitions of terms used in the text of this report, and Appendix C provides a list of acronyms included in this report.

2.0 APPLICABILITY OF CHEMICAL ABSTRACTS SERVICE REGISTRY NUMBERS IN THE EPA ENVIRONMENT

This section addresses the origin of the CASRN and its applicability to identification of chemical substances in the EPA regulatory environment.

2.1 Background for CAS Registry Numbers

CASRNs are unique identification numbers, with no inherent meaning, assigned by CAS to chemical substances when they are identified in scientific literature that is reviewed by CAS indexers and abstractors. In addition to assigning CASRNs to specific, definable substances, CAS has also assigned CASRNs to chemical substances that are not specific (e.g., mixed isomers, tars and waxes, substances "not otherwise specified" (NOS)) and to specific ions and isotopes when those chemical substances were referenced in published materials. Several million CASRNs have been assigned to date.

The CASRN is used nationally and internationally to identify chemicals in information systems that contain chemically-related data. Although CAS has indicated a willingness to assign Registry Numbers to non-specific substances when registry numbers are needed for regulatory purposes, CASRNs may not be available at this time to identify all chemical substances of interest to EPA's regulatory concerns.

2.2 Issues Related to the EPA Chemical Data Standard

The EPA chemical data standard specifies that chemical data elements other than CASRN are optional for identifying a chemical substance in a system implementation. The registry number alone, however is not sufficient to identify a chemical substance to a system user for reasons which include the following:

- CASRNs have no inherent meaning to describe the composition and structure of the chemical or to enable grouping of related chemicals.

- System users will rarely know the CASRN of the chemical of interest and will be unable to conduct searches based on that number alone.

- The use of CASRN alone does not provide a means for verifying the identity of the chemical. This is of particular concern where the CASRN has been miskeyed or is used inaccurately.

2.3 Characteristics of Chemicals Monitored by the EPA

The characteristics of chemicals regulated and monitored by the EPA cause concerns relevant to the implementation of the CASRN within the EPA environment. Characteristics of EPA regulated chemicals that raise data management concerns include the following:

- The EPA regulates chemical substances that are not specific, definable chemicals (e.g., the chemical categories regulated under Title III of the Emergency Planning and Community Right-to-Know Act (EPCRA) and the hazardous wastes regulated by the Resource Conservation and Recovery Act (RCRA). Categories regulated by EPCRA Section 313 are identified in TRIS by category codes (e.g., Nnnn); RCRA hazardous wastes are identified by hazardous waste codes (e.g., Fnnn and Knnn). CASRNs do not exist for many of these categories.

- EPA environmental systems require and maintain information for groups of chemical substances (e.g., 2,4-D salts and esters) as well as for individual substances that are components of the groups (e.g., 2,4-D tert-butyl ester).

- Chemical substances are sometimes monitored as ions (e.g., Nitrates or Cyanides) and sometimes as salts (e.g., Sodium Nitrate or Copper Cyanide).

- Different radioisotopes of a substance are monitored (e.g., Iron^{55} and Iron^{59}) in addition to different valence states of the element (e.g., Ferrous Chloride and Ferric Chloride) as well as the element in its natural state (Iron, in this example).

- Some reported chemicals are claimed to be Confidential Business Information (CBI) or Trade Secret by the regulated community. The identity of the chemical in this context must be restricted to authorized users and not available for general access.

- New chemicals reported under Section 5 of the Toxic Substances Control Act (TSCA) may not have been published and consequently not have had CASRNs assigned.

2.4 Concerns for Integrating Chemicals Monitored by EPA

Concerns for linking chemicals in EPA systems have been documented in the *Audit Analysis Report for the Data Standards Implementation Program* and in the *Mission Needs Analysis for* *a Central Chemical Index System*. The following is a brief review of those concerns. Appendix B provides definitions of terms used in this subsection.

- The CASRN has no capability to link different forms of a chemical substance or to group related chemicals. Therefore, the use of CASRNs does not address the need to report and integrate data on groups (e.g., categories and mixtures) and other variations (e.g., isomers, isotopes, and ionic materials).

- Different systems often monitor different variations of the same substance (e.g., stereoisomers), each of which can be represented by a different CASRN. An example of a substance monitored as more than one stereoisomer in different EPA systems is hexachlorocyclohexane, also known as Benzene Hexachloride (BHC), which has the following CASRNs:

Isomer	**CAS Number**	**Monitored by**
BHC (unspecified)	608-73-1	PCS (parm = 81283)
BHC (alpha)	319-84-6	CERCLIS
BHC (beta)	319-85-7	CERCLIS
BHC (delta)	319-86-8	CERCLIS
BHC (gamma) - Lindane	58-89-9	CERCLIS, RCRIS, TRIS, FRDS

  The absence of a mechanism to identify and group related information across systems makes it difficult to integrate chemical data.

- Inappropriate use of CASRNs in EPA systems complicates the integration of data across systems. Some systems use a CASRN to denote categories, even though the number refers to a specific chemical. For example, RCRIS uses CASRN 94-75-7 (the CASRN for 2,4-D acid) to represent 2,4-D salts and esters; CERCLIS uses the same number to represent 2,4-D esters. In addition, the category "Polybrominated biphenyls, NOS" is represented in EMMI with CASRN 59536-65-1, a CASRN identified in the *Dictionary of Chemical Names and Synonyms* as "Hexabromobiphenyl," one of many Polybrominated biphenyls. TRIS identifies "Polybrominated biphenyls" with a category code.

- CASRNs in EPA systems have not been entered with consistent formats. Variations include right or left justification, with or without hyphens or leading zeros (e.g., formaldehyde may be represented as 000050000, 50000 or 50-00-0). These variations impede automated matching of CASRNs.

- EPA systems do not validate the CASRN with the CAS check digit, resulting in data entry errors that preclude data integration.

3.0 POSSIBLE DATA ELEMENTS FOR CHEMICAL DATA REPRESENTATION

Data types by which chemical data are traditionally represented in both manual and automated systems to identify and index chemical substances include the following:

- **Registry Number** -- A registry number (i.e., chemical identification code) is necessary in automated systems to link the identity of a chemical substance to data relevant to that substance. As stated previously in Section 2.2, a registry number alone is inadequate to satisfy users' needs for identifying a chemical substance. All EPA program office systems use a chemical identification code to link the identify of regulated substances to relevant environmental data. Examples of the codes used by ENVIROFACTS component systems are listed as follows:

System	Code
PCS	Parameter codes in the field PARAMETER_CODE. The codes represent multiple parameters, resulting in more than one code for the same substance.
TRIS	CASRNs and category codes in the field TRI-CHEMICAL-ID.
CERCLIS	CASRNs and category codes in the field C3701.
RCRIS	Hazardous Waste Codes in the Handler2 file; CASRNs in the Corrective Action file.

- **Chemical Name** -- Names or descriptions are used in all EPA environmental systems to describe chemical substances. A chemical name alone is not an appropriate data item to integrate chemical data across EPA data systems for reasons that include the following:

  - One chemical has many different names, including systematic names, index names, common names, synonyms, and trade names (i.e., names of commercial products). Individual EPA systems often use different names to refer to the same substance.

  - Chemical names are sometimes misleading. For example, Hexachlorobenzene (CASRN 118-74-1) is not the same chemical as Benzene Hexachloride (Hexachlorocyclohexane), but the common name can be misinterpreted.

  - Slight differences in spelling of a chemical name can lead to a complete misrepresentation of the chemical substance. For example, Ethanol is an alcohol found in alcoholic beverages; Ethanal is an aldehyde that is not intended for human ingestion.

- **Molecular Formula** -- Molecular formula is an indexing tool that lists the atomic composition of a molecule (i.e., the type and number of atoms). Many different chemical substances have the same molecular formula; therefore, molecular formula is inadequate to identify and integrate chemical substances across EPA systems. For example, chemicals represented by C_2H_6O include both Ethanol and Dimethyl Ether. The Office of Pollution Prevention and Toxics (OPPT) is the only EPA program office that stores molecular formula in its environmental data systems. The molecular formula is not conveniently available to capture for the EMCI through automated processes.

- **Structural Formula** -- Structural formula is a modification of the molecular formula where the atom types and number are listed in a format to provide information about their connectivity. For example, CH_3OCH_3 is the structural formula for Dimethyl Ether, and C_2H_5OH is the structural formula for Ethanol. There is no automated source of this data element for the EMCI, since no EPA environmental systems currently store the structural formula. This data element is therefore inappropriate for the prototype EMCI.

- **Connectivity Tables** -- Connectivity tables are tables which record the exact atoms and their connectivity within a molecule. The tables enable exact chemical structural information to be stored, structure and substructure to be searched, and two-dimensional structure representations to be displayed. OPPT, the only EPA program office to use connectivity tables, uses this functionality to identify related chemicals for risk assessment of new chemicals. Analysis of molecular structure is not required in a prototype EMCI. The data elements used by program office systems for chemical identification and which are readily available for automated update to the EMCI are limited to a non-standard chemical identification number and a non-standard chemical name. These data elements serve as the basis for defining data elements for the EMCI.

4.0 PROPOSED RULES FOR CHEMICAL DATA REPRESENTATION

This section presents the proposed rules for representing chemical data in the EMCI. It includes the scope of chemicals for which the rules are applicable and the proposed rules for chemical representation.

4.1 Scope of Rules for Chemical Data Representation

These proposed rules for implementation of the CASRN Data Standard in the GATEWAY/ENVIROFACTS environment shall be applicable for representing the following kinds of chemical data:

- Specific, unique chemical substances (i.e., chemicals for which the exact chemical structure is known).

- Chemical substances with undefined molecular structure (e.g., Paraffin Wax or Tall Oil).
- Chemical substances for which CASRNs have not been assigned.

- Ions, where the complete salt is not specified (e.g., Ammonium, Nitrates, Chlorates, etc.).

- Categories and/or groups of chemicals, including the following:

  - Salts and/or esters of a particular acid.

  - Salts of a particular metal (e.g., Mercury).

  - Mixtures [e.g., Cupric Oxide (CuO) mixture with Cuprous Oxide (Cu_2O), or Cresol (mixed isomers)].

  - Chemical substances with different:

    - Number of substituents (e.g., Chlorinated Phenols).
    - Positional isomers (e.g., Dichlorophenol).
    - Valence states (e.g., Ferric and Ferrous Iron compounds).
    - Stereoisomers (e.g., alpha, beta, delta and gamma BHC).
    - Radioisotopes (e.g., $Iron^{55}$ and $Iron^{59}$).

- Arbitrary groupings of chemicals (e.g., chemicals that appear on a regulatory list).

- Chemicals that have been claimed CBI or Trade Secret in unsanitized systems. Sanitized systems will conceal the chemical identity of chemicals that have been claimed CBI or Trade Secret.

4.2 Proposed Rules for EMCI Chemical Data Representation

Of the traditional chemical data types described in Section 3.0, the only chemical data used across EPA environmental systems are some type of identification number and some kind of name or description of the chemical substances regulated or monitored by each. These data form the basis for the EMCI design. The EMCI will require storage of the following data for each chemical substance:

- Identification number.
- The most specific, systematic chemical name available.

Other names and synonyms (i.e., names used in the ENVIROFACTS component systems) will also be stored.

4.2.1 Identification Number

Chemical substances shall be identified by a CASRN wherever possible, including mixtures, categories, ions, radioisotopes, and ambiguous substances. Where CASRNs are not available, the EMCI software shall assign a unique identification number to be used for identification in EMCI. Identification numbers assigned by EMCI shall be prefixed with a "U" to indicate that the CASRN is unknown.

The following rules shall be applied to the identification of chemical substances:

- Specific, unique chemicals shall be identified with a CASRN.

  - The number shall pass check digit validation at data entry.

  - The number shall have been previously assigned by CAS for this specific, unique chemical. *It is not intended that CAS be requested to assign CASRNs for EPA data entry.*

  - The number shall be stored electronically as a right-justified, 9-digit number with leading zeros.

    Example: 001897456

  - CASRNs shall be displayed to conform with the CAS edit format of "nnnnnn-nn-n."

    Example: 1897-45-6

- Non-specific, ambiguous chemicals shall be identified with a CASRN, if available.

  Example: Tallow -- CASRN 61789-97-7

  Where CASRNs are not readily available from the submission forms, published references (e.g., the *Merck Index* or the *Dictionary of Chemical Substances*) and existing EPA chemical data systems shall be consulted to determine if a CASRN exists for that substance. Where CASRNs are not available, the EMCI software shall assign a unique identification number to be used for identification in EMCI.

  Example: Cinnamon -- U00000059

- Mixtures and categories of chemicals shall be identified with a CASRN, if available.

  Example: Polychlorinated Biphenyls -- CASRN 1336-36-3
  Dinitrotoluene (mixed isomers) -- CASRN 25321-14-6

  Where CASRNs are not available, the EMCI software shall assign a unique identification number to be used for identification in EMCI.

  Example: 2,4,5-TP, salts and esters -- U00000005
  DDT and Metabolites -- U00000071

  Note: Rules for specifying categories and mixtures in EMCI are described in Section 4.3 of this report.

- Ionic substances shall be identified with a CASRN, if available.

  Example: Cyanide -- CASRN 57-12-5
  Nitrate -- CASRN 14797-55-8

  Where CASRNs are not readily available, the EMCI software shall assign a unique identification number to be used for identification in EMCI.

  Example: Nitrates/Nitrites -- U00000262

- Radioisotopes shall be identified with a CASRN, if available.

  Example: Thorium^{230} -- CASRN 14269-63-7

  Where CASRNs are not readily available, the EMCI software shall assign a unique identification number to be used for identification in EMCI.

- Only correct CASRNs will be used in the Master List of chemicals in the EMCI. The CASRNs shall be verified with the numbers used in the RoL. A *correct CASRN* is one that has been assigned by CAS to a specific, definable chemical substance or group of substances and that conforms to the CAS check digit algorithm.

  Where an *incorrect CASRN* has been entered into an ENVIROFACTS component system with a typographical error or where the CASRN has been applied to a chemical substance or group of substances different from that to which it was assigned by CAS, the EMCI will cross reference the chemical substance to its correct CASRN and the EMCI data management team shall notify the program office of the error in the program office system.

- Correct CASRNs and EMCI-assigned identification codes shall be cross-referenced to the system identification codes used by the ENVIROFACTS component systems.

4.2.2 Chemical Names

A **specific, unique name** is required for each chemical substance to provide an unambiguous definition of the substance.

- Specific, unique chemicals shall be identified with a systematic name (e.g., a CAS index name or IUPAC name). The systematic name shall include:

  - Positional indicators (e.g., "2,4-" or "1,1'-").

  - Stereo indicators (e.g., alpha, beta, gamma, etc.; cis or trans; L or D; etc.).

  - Radioisotope identification (e.g., Thorium^{230}, Thorium^{232}; Uranium^{235}, Uranium^{238}; etc.).

  - Valence state (e.g., Ferrous or Ferric).

  Where available, EMCI will use the same systematic name as is used by the RoL for its master list of chemical identities. The RoL uses the CAS index name as obtained from TSCA Inventory lists.

- The systematic name shall be entered in inverted format, to facilitate grouping of related chemicals.

  Example: 1,3-Benzenedicarbonitrile, 2,4,5,6-Tetrachloro-

- For ambiguous chemical substances, chemical names shall be as specific as possible to clearly indicate the characteristics of the substance.

  Example: Diesel Fuel, Marine

- For non-specific substances, a suffix will be added to indicate the nature of the ambiguity. For example:

- Mixed isomers -- to indicate that the substance is known to contain a mixture of isomers.

  Example: Dichlorobenzene, Mixed Isomers

- NOS -- to indicate that the composition of the substance was not specified.

  Example: Copper and Compounds, NOS

Synonyms for the chemical (i.e., the chemical names, including common or trivial names that are used in the ENVIROFACTS component systems) shall also be included in EMCI.

Examples: Chlorothalonil
Tetrachloroisophthalonitrile
1,3-Dicyano-2,4,5,6-tetrachlorobenzene

- Trade names may be entered, where the regulated or monitored substance is the active ingredient of that commercial product. Only trade names that are stored in the ENVIROFACTS component systems shall be included in the EMCI.

  Example: Banvel (a trade name for Dicamba)

- For identification of ionic substances, the chemical name in program office systems shall be as specific as possible to clearly indicate the material being monitored or regulated (e.g., Inorganic Nitrates, Aqueous Ammonium, etc.).

- Synonyms derived from the program office systems' records may also include, as a suffix, the process information or physical state of a substance where that information is stored in the program office system record. Examples of suffixes to be used include:

  - Dust.
  - Fume.
  - Friable.
  - Liquid.
  - Vapor.
  - Aqueous solution.
  - Manufacturing.
  - Strong-acid process.
  - Fibrous form.

4.3 Proposed Rules for Grouping Chemicals in EMCI

The rules listed in this section shall be applied for the specification of categories, mixtures, and substances that are "not otherwise specified" (NOS). The purpose of the groupings is to enable the groups to be related to the specific chemicals that are components of the categories, mixtures and NOS chemical substances. Refer to Appendix B for definitions of terms used in this subsection.

The following proposed rules shall be applied to the grouping and identification of categories, mixtures and NOS substances.

- Categories, mixtures, and non-specific substances shall be identified in the EMCI as categories, mixtures, or "not otherwise specified." These substances shall be cross referenced to other specific chemicals that might be contained in that grouping.

- Categories, mixtures, and non-specific substances will be identified by CASRNs and CAS systematic name where those names and numbers exist. For example: "1,1'-Biphenyl, Chloro Derivs," shall be identified by CASRN 1336-36-3.

- Identification numbers shall be assigned by the EMCI software for categories, mixtures, and NOS substances that cannot be identified by a CASRN. These shall be prefixed with a "U" to indicate that the CASRN is unknown.

- Hazardous Waste Codes for which no CASRN has been assigned will be identified in EMCI by an EMCI identification number (prefixed with "U" and assigned by the EMCI software). The waste code will be stored as a prefix to the descriptive name of the hazardous waste.

- Where there is no CASRN or CAS name for a mixture or category, the name on the master list shall be as specific as possible to describe the exact nature of the mixture or category.

- Each chemical substance on the master list must be unique, including specific chemicals, mixtures, and categories.

- Categories of chemical substances shall be grouped by chemical relationships, not by regulatory concerns. For example, the category "Chlorophenols" shall include all chlorinated phenols regulated by EPA program office systems. Separate categories of chlorinated phenols relevant to SARA 313 or to CWA 307(A) will not be maintained in EMCI. For lists of chemicals and categories of chemical substances relevant to a regulation, the user shall refer to the RoL, which maintains that information.

  Note: The issue of chemical data integration by regulation, needs to be addressed by a separate effort.

5.0 RULES FOR CHEMICAL DATA MANAGEMENT

Chemical data management for the EMCI involves those who provide the data, those who manage the data in the ENVIROFACTS environment, and the organization of that data in ENVIROFACTS.

5.1 Responsibilities

Responsibilities for chemical data in the EMCI are shared by the program offices and by the ENVIROFACTS support teams.

5.1.1 Program Offices

Program offices for component systems in the ENVIROFACTS system will have the following responsibilities for the EMCI:

- Provide up-to-date documentation, including Data Element Dictionaries and data administration procedures, for use by the ENVIROFACTS team to ensure their complete understanding of the data and data relationships in the component systems.

- Provide the ENVIROFACTS team with access to the program office system for regularly scheduled downloading of the environmental data and the chemical data validation tables.

- Accept and respond to data inconsistencies identified by the EMCI Chemical Data Management Team.

5.1.2 ENVIROFACTS Chemical Data Management Team

The chemical data management team assigned to ENVIROFACTS shall be responsible for overseeing the data quality of the EMCI. The following are representative of the team's responsibilities:

- Maintain the chemical data files in EMCI as they are acquired from the component systems. The team shall:

  - Determine additions and changes to the chemical data validation tables received from the component systems.

  - Update the program office systems' chemical data tables in EMCI to conform to changes made to those tables by the program office system.

- Update the EMCI master chemical list and cross references to reflect changes to the component systems. Any uniquely defined chemical substance, whether single chemical, mixture, category or NOS, shall be represented by only one entry on the EMCI master chemical list.

- Determine the correct CASRN to be used in EMCI for identification of a chemical substance.

- Determine the systematic name to be used in the EMCI master chemical list for identification of a chemical substance.

- Identify and group categories and mixtures into a hierarchical system consistent with groups of chemicals identified in the component ENVIROFACTS systems.

  Note: Where appropriate, chemicals in the EMCI will be grouped according to the same rules as they are grouped in the RoL.

- Assign identification numbers for any chemical substance or group of substances, where a CASRN does not exist.

- Correct chemical names and CASRNs where errors are noted in the master chemical list, and maintain accurate cross references with the ENVIROFACTS component systems.

- Notify the system managers of component systems when errors or inconsistencies are observed by the data management team in the use of CASRN or other codes in component systems.

5.2 Data Organization

The following rules shall be applied to the management and organization of chemical data in the EMCI:

- EMCI data shall be organized consistently with other data managed in ENVIROFACTS to ensure maximum integration of chemical data in ENVIROFACTS (e.g., file names and data attribute names shall conform to established standards).

- Access to EMCI shall be consistent with access methods currently used for the ENVIROFACTS system.

- A Master List of chemical substances shall be maintained, containing a reference to all chemicals included in the program office systems that are components of ENVIROFACTS.

- Lists of chemical substances relevant to each component program office system shall be maintained with the same descriptive data (e.g., chemical identification code and chemical name) as they appear in each component system.

- Lists of chemicals contained in component systems shall be identified according to the chemical identification codes used in each component system.

- Each chemical shall be identified by the component computer systems that store environmental data for the chemical.

- Cross references shall be developed to link the CASRN used in the EMCI to the identification code used by the component program office system.

- The regulations which apply to a chemical are available through the RoL and are not planned for duplication in the EMCI. The need to integrate chemical data based on regulations will be addressed as a separate effort.

6.0 CONCLUSIONS

The EPA has issued a data standard for chemical data representation, requiring that the CASRN be used in all EPA computer systems to identify specific, unique chemical substances. The EPA, however, regulates and monitors chemical substances for which CASRNs may not exist (e.g., non-specific chemicals, mixtures, categories, ions, stereoisomers, and radioisotopes). The EPA system users need to search for substances that are chemically related (e.g., 2,4,5-T esters and salts), in addition to searches for specific chemicals. They also need to search for chemicals across program office systems and media (e.g., air, water, etc.).

All program offices have not implemented the chemical data standard, and those that have implemented it do not use the CASRN in a consistent manner that enables linkages across systems. Therefore, a chemical integrator module will be developed for the ENVIROFACTS integrated system to serve as a model for integration of chemical data across all EPA environmental systems. The system will also facilitate implementation of the chemical data standard.

This document includes proposed rules for chemical data representation and chemical data management for a master chemical integrator. The proposed rules will be used for data representation and management in the EMCI and serve as a preliminary guide towards enhancing the EPA chemical data standard.

The EPA intends this document to be a living document that will be revised and appended through out the development of the EMCI to conform to data needs, system constraints and the users' needs to search and retrieve data. This document will be further enhanced as needed to meet the requirements for the future development of an Agency-wide Chemical Index System.

APPENDIX A

References

References used in the preparation of this document include the following:

- Chemical Abstracts Service Registry Number Data Standard, EPA 2180.1, June 26, 1987.

- Audit Analysis Report for the Data Standards Implementation Program, SDC-055-013-2002A, January 8, 1993.

- Recommendations for a Data Standards Implementation Strategy, SDC-0055-013-AM-2006A, February 26, 1993.

- Mission Needs Analysis for a Central Chemical Index System, SDC-0055-013-AM-2007A, March 15, 1993.

- GATEWAY/ENVIROFACTS Database Description and Maintenance Procedures, SDC-0055-051-ER-2007, December 17, 1993.

- TRIS Physical Design (Mainframe), prepared by SYCOM, Inc. for the Office of Pollution Prevention and Toxics (OPPT), November 9, 1992.

- TRIS Reporting Form R, EPA Form 9350-1 (Rev. 5/14/92).

- The Register of Lists (RoL), developed by the Office of Policy, Planning and Evaluation, Office of Regulatory Management & Evaluation, Information Policy Branch, Fall 1992.

- Environmental Monitoring Methods Index (EMMI), Version 1.0 Software, Software and User's Manual, December 1991.

- Proposed Rules for Specifying Categories and Mixtures in the Register of Lists, USEPA, Region 5, December 3, 1993.

- Register of Lists: Data Discrepancies Between CERCLA List and Constituent Lists, USEPA, Region 5, December 3, 1993.

- Dictionary of Chemical Names and Synonyms, P. H. Howard and M. Neal, Lewis Publishers, 1992.

APPENDIX B

Definition of Terms

The following are definitions of terms used in this document for proposed rules for identification of chemical data:

- **Chemical Abstracts Service Registry Number (CASRN)** -- A unique, identifying number assigned by CAS to each distinct chemical substance recorded in the CAS Chemical Registry System.

- **CAS Index Name** -- The systematic name recorded in the CAS Chemical Registry System for each registered substance, presented in an inverted format to facilitate arrangement of related chemicals in an organized index.

- **Molecular Formula** -- A formula that lists the kind and number of atoms in a molecule.

- **Structural Formula** -- A formula that lists the kind and number of atoms in a molecule in a format that represents the molecular structure.

- Chemical **Category** -- A substance name that encompasses a class or group of specific chemicals. For example, the category "chlorophenols" refers to chlorinated phenols that contain one or more chlorine atoms (e.g., 2-Chlorophenol, 2,4-Dichlorophenol, Pentachlorophenol, and others).

- **Not Otherwise Specified** (NOS) -- A chemical substance for which the exact position of substituents or the spatial configuration is not specified. For example:

  - Dichlorobenzene (CASRN 25321-22-6) might represent 1,2-Dichlorobenzene (CASRN 95-50-1), 1,3-Dichlorobenzene (CASRN 541-73-1), and 1,4-Dichlorobenzene (CASRN 106-46-7).

  - 2-Butene (CASRN 107-01-7) might represent 2-Butene-trans (CASRN 624-64-6) or 2-Butene-cis (CASRN 590-18-1).

- Chemical **Mixture** -- One of the following types of chemical substance:

  - More than one isomer of a particular chemical (e.g., "Cresol, mixed isomers").

  - More than one unique chemical (e.g., "Copper Chloride Hydroxide ($Cu_2Cl(OH)_3$), mixture with Copper Hydroxide Sulfate ($Cu_4(OH)_6(SO_4)$")

  Note: A mixture can be reported because the mixture itself is regulated (e.g., Hazardous Waste Code F001, spent halogenated solvents used in degreasing) or because the regulated chemical is a component of a mixture and must therefore be reported (e.g., as required under EPCRA Section 313).

- **Ambiguous Chemicals** -- Substances of a chemical nature that cannot be defined by a specific molecular formula or structure (e.g., Coal Tar, Lanolin Wax, Tallow, Tall Oil, etc.).

- **List** of Regulated Chemicals -- A group of chemicals that are related only because they are included in the same regulation.

- **Systematic Name** -- A systematic chemical name is one that describes structural detail, including:

  - Positional indicators (e.g., "2,4-" or "1,1'-").

  - Stereo indicators (e.g., alpha, beta, gamma, etc.; cis or trans; L or D; etc.).

  - Radioisotope identification (e.g., Thorium^{230}, Thorium^{232}; Uranium^{235}, Uranium^{238}, etc.).

  - Valence state (e.g., Mercurous or Mercuric; Chromium (III) or Chromium (VI)).

- **Isomer** -- A compound with the same percentage composition and molecular weight as another compound but differing in chemical or physical properties. Such a compound may differ based on manner of:

  - Linkage of constituent atoms.

  - Arrangement of its constituent atoms in space.

  - Rotation of polarization of light to right or left.

  - Structural asymmetry about a double bond in the molecule.

- **Ions** -- An atom or group of atoms that has acquired or is considered to have acquired a net electric charge by gaining electrons in or losing electrons from an initially electrically neutral configuration.

- **Isotope** -- One of two or more atoms whose nuclei have the same number of protons but different numbers of neutrons.

- **Radioisotopes** -- A naturally occurring or artificially produced radioactive isotope of an element.

APPENDIX C

Acronyms

The following is a list of acronyms used in this document:

CAS	Chemical Abstracts Service
CASRN	Chemical Abstracts Service Registry Number
CBI	Confidential Business Information
CERCLA	Comprehensive Emergency Response, Compensation and Liability Act
CERCLIS	CERCLA Information System
EMCI	ENVIROFACTS Master Chemical Integrator
EPA	Environmental Protection Agency
EPCRA	Emergency Planning and Community Right-to-Know Act
EMMI	Environmental Monitoring Methods Index
FINDS	Facility Index System
GLNPO	Great Lakes National Program Office
NOS	Not Otherwise Specified
OIRM	Office of Information Resource Management
OPPE	Office of Policy, Planning and Evaluation
OPPT	Office of Pollution Prevention and Toxics
OWRS	Office of Water Regulations and Standards
PCS	Permit Compliance System
RCRA	Resource Conservation and Recovery Act
RCRIS	RCRA Information System
RoL	Register of Lists
SARA	Superfund Amendment Reauthorization Act
TRIS	Toxic Release Inventory System
TSCA	Toxic Substances Control Act